

Collaboration Activities in the Geosciences Network (GEON)

Chaitan Baru

PI, GEON

Director, Science R&D

San Diego Supercomputer Center

Outline

- **About GEON**
- **Collaboration “modes” in GEON**
 - Barriers / incentives for collaboration
- **The SDSC/Calit2 Synthesis Center**
 - The SDSC Notebook project

About *GEON*

- **National Science Foundation ITR Project, 2002-2007, \$11.6M**

PI Institutions

- **Arizona State University**
- **Bryn Mawr College**
- **Penn State University**
- **Rice University**
- **San Diego State University**
- **San Diego Supercomputer Center/UCSD**
- **University of Arizona**
- **University of Idaho**
- **University of Missouri, Columbia**
- **University of Texas at El Paso**
- **University of Utah**
- **Virginia Tech**
- **UNAVCO**
- **Digital Library for Earth System Education (DLESE)**

Partners

- **California Institute for Telecommunications and Information Technology, Cal-(IT)²**
- **Chronos**
- **CUAHSI-HIS**
- **ESRI**
- **Geological Survey of Canada (GSC)**
- **HP**
- **IBM**
- **IRIS**
- **Kansas Geological Survey**
- **Lawrence Livermore National Laboratory**
- **NASA Goddard, Earth System Division**
- **Southern California Earthquake Consortium (SCEC)**
- **U.S. Geological Survey (USGS)**

Affiliated Project

- **EarthScope**

Project Goals and Approach

- **Develop *cyberinfrastructure* to support the “day-to-day” conduct of science (*e-science*), not just “hero” computations**
 - Based on a Web/Grid services-based distributed environment
- **Work closely with geoscientists to help create data sharing frameworks, best practices, and useful and usable capabilities and tools**
- **The “two-tier” approach**
 - Use best practices, including commercial tools,
 - while developing advanced technology in open source, and doing CS research
- **Leverage from other intersecting projects, e.g. BIRN, SEEK, OptIPuter**

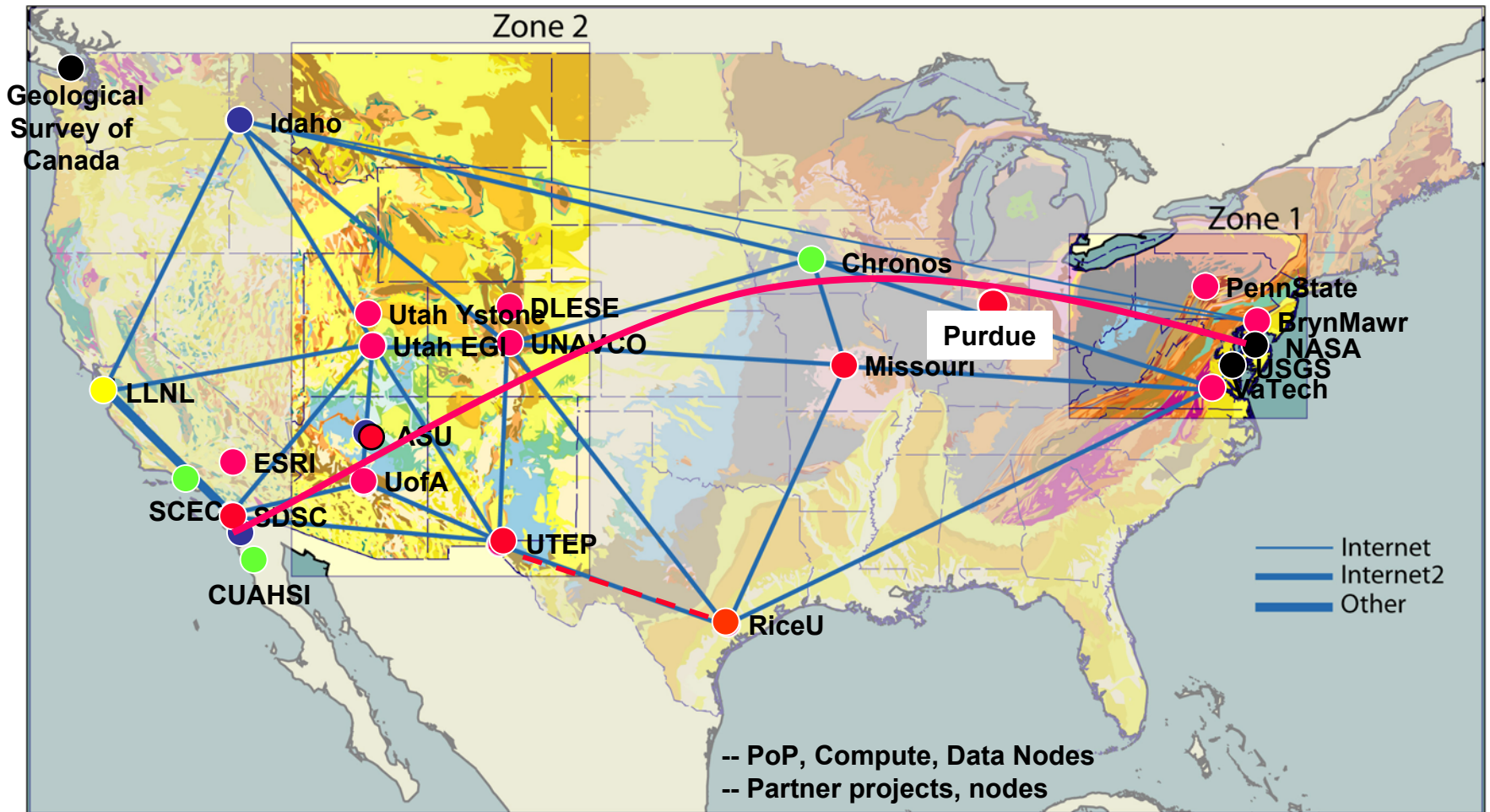
A Brief History of Collaboration in GEON

- **Began with “Geoinformatics” workshops sponsored by NSF**
 - First NSF Geoinformatics workshop, October 1999
 - Second workshop, April 2000
 - Only “domain” scientists. No involvement of IT researchers.
- **NSF made the introductions between SDSC personnel and Geoinformatics organizers**
- **Third workshop, September 2000, was attended by SDSC**
- **Visit to SDSC by key Geoinformatics PI’s**
 - Identified the key IT research issues: sophisticated data integration, distributed/grid computing, 4D and higher-order data visualization
- **Project funded under NSF ITR program in 2002**
 - ...collaborative science is underway
- **Several collaborations are emerging with other geoscience and other sciences projects, international partners, and the EarthScope project in the US.**

Large Projects and Collaboration

- **At some level, all large projects are about collaboration.**
 - Sounds self-evident, but we never really seem to plan from the beginning for this
- **Collaboration for science versus collaboration to develop tools**
 - Is collaboration an inherent problem when the “tool is the end goal”...
 - Rather than the science being the end goal?

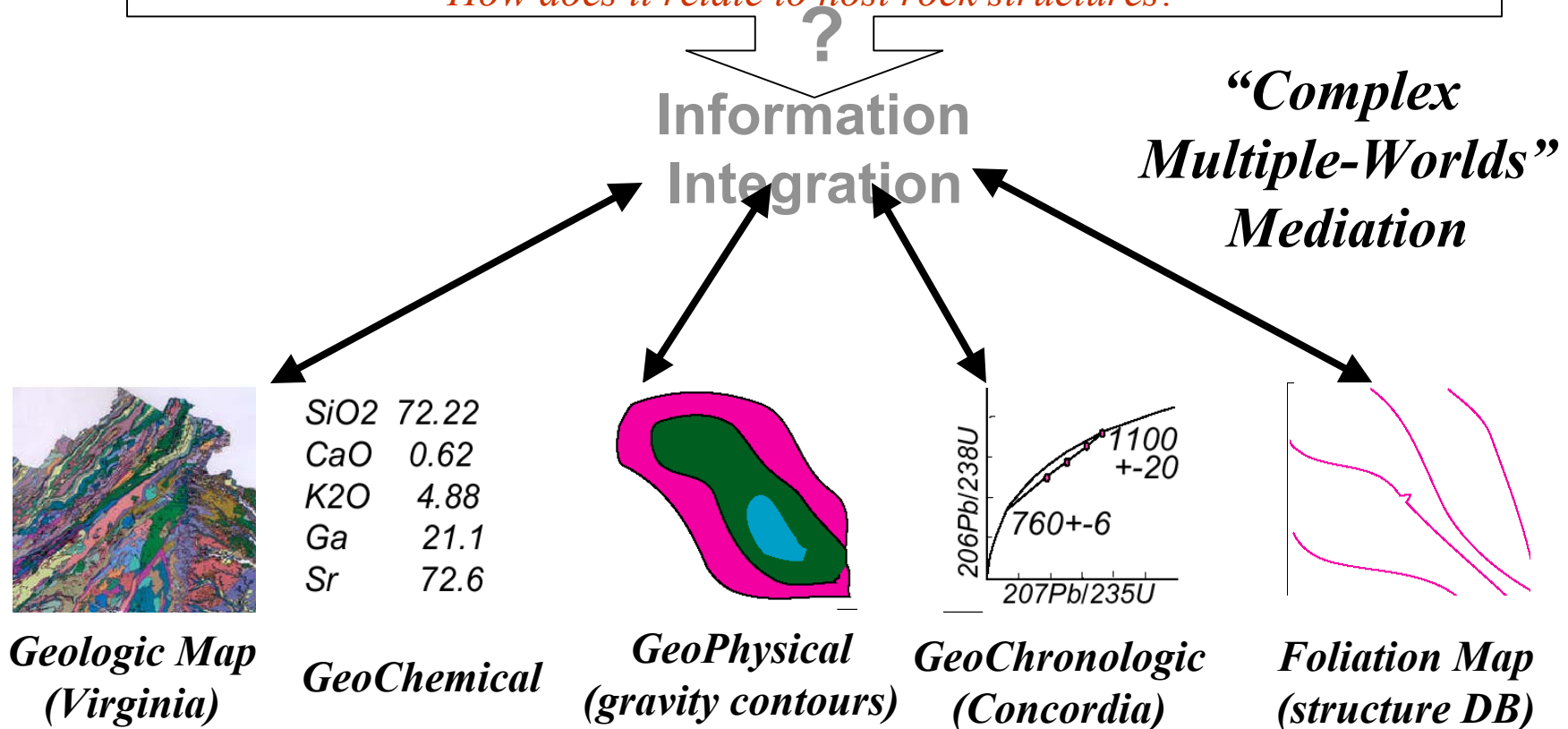
The GEONgrid



The Need to Collaborate: Integration of multi-disciplinary data sets

Example:

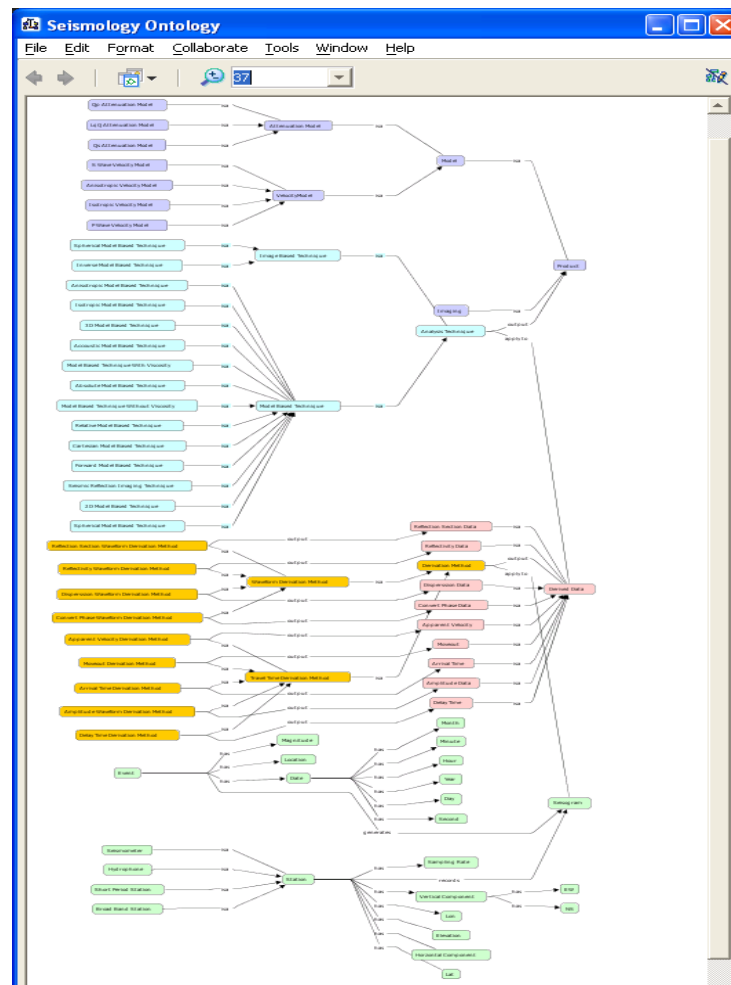
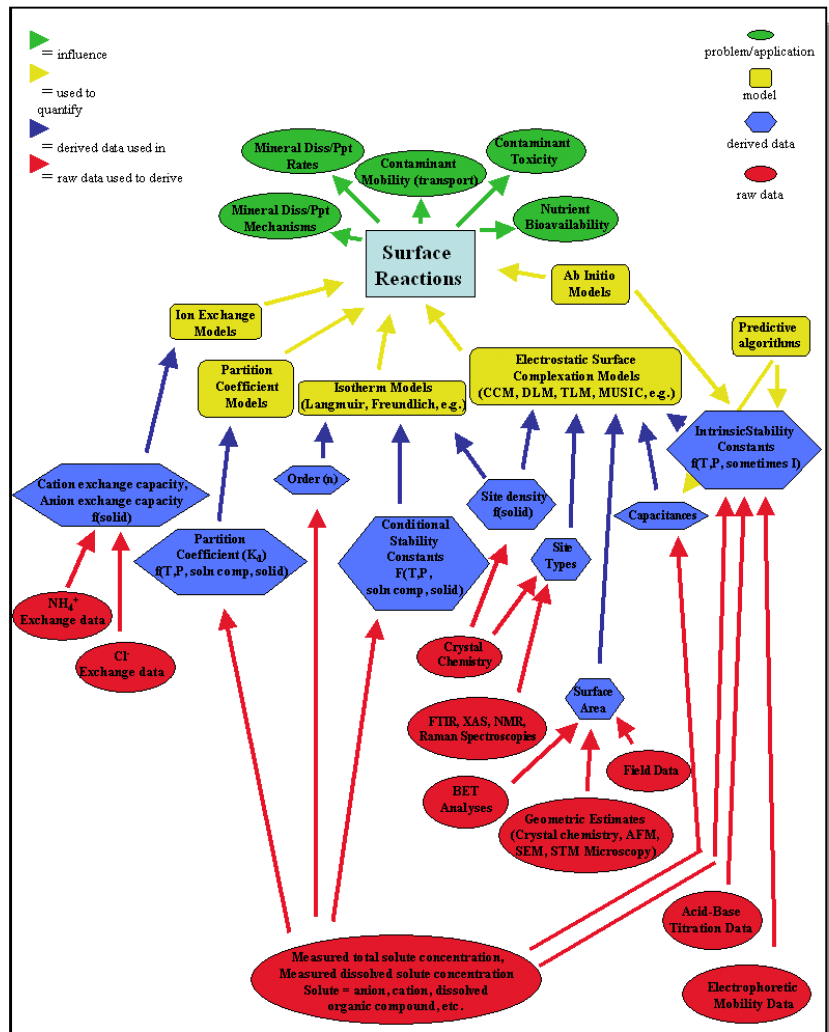
*What is the distribution and U/ Pb zircon ages of A-type plutons in VA?
How about their 3-D geometry?
How does it relate to host rock structures?*



Development of Shared Knowledge Structures

- ***Conceptual models*** of a domain or application, for purposes of communication, and/or system design
- **Classification of ...**
 - concepts (taxonomy) and
 - data/object instances through classes
- **Analysis of ontologies e.g.**
 - **Graph queries** (reachability, path queries, ...)
 - **Reasoning** (concept subsumption, consistency checking, ...)
- **Targets for semantic data registration**
- **Conceptual indexes and views for**
 - searching,
 - browsing,
 - querying, and
 - integration of registered data

Creating and Sharing Concept Maps



**Bill Glassley (LLNL), Randy Keller (UTEP),
Bertram Ludaescher, Kai Lin,
Dogan Seber (SDSC), et al**

Community-Based Ontology Development

- **Focused meetings**
- **Bring scientists together for 2+ days**
- **Include participation by Computer Science / Knowledge Base Management experts**
- **Create concept maps**
- **Refine**
- **Iterate**
 - ➔ from napkin drawings, to concept maps, to ontologies
- **Need better, online collaboration tools for this**


...implemented as a web portal

HERO | codex - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Search Favorites Media Mail

Address <http://hero.geog.psu.edu/codex/jsp/workspace/Home.jsp> Go

 **codex**

Bill Pike [\[Sign out\]](#) HERO Project

Bill Pike's workspace

[Edit personal settings](#)
[Send a message](#)
[Help](#)

[Create a new concept...](#)

Concepts

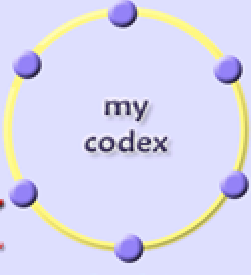
Tasks
[Start a new task...](#)

Files
[Add a new file...](#)



Tools
[Add a new tool...](#)

Places
[Add a new place...](#)

Groups
[Create a new group...](#)

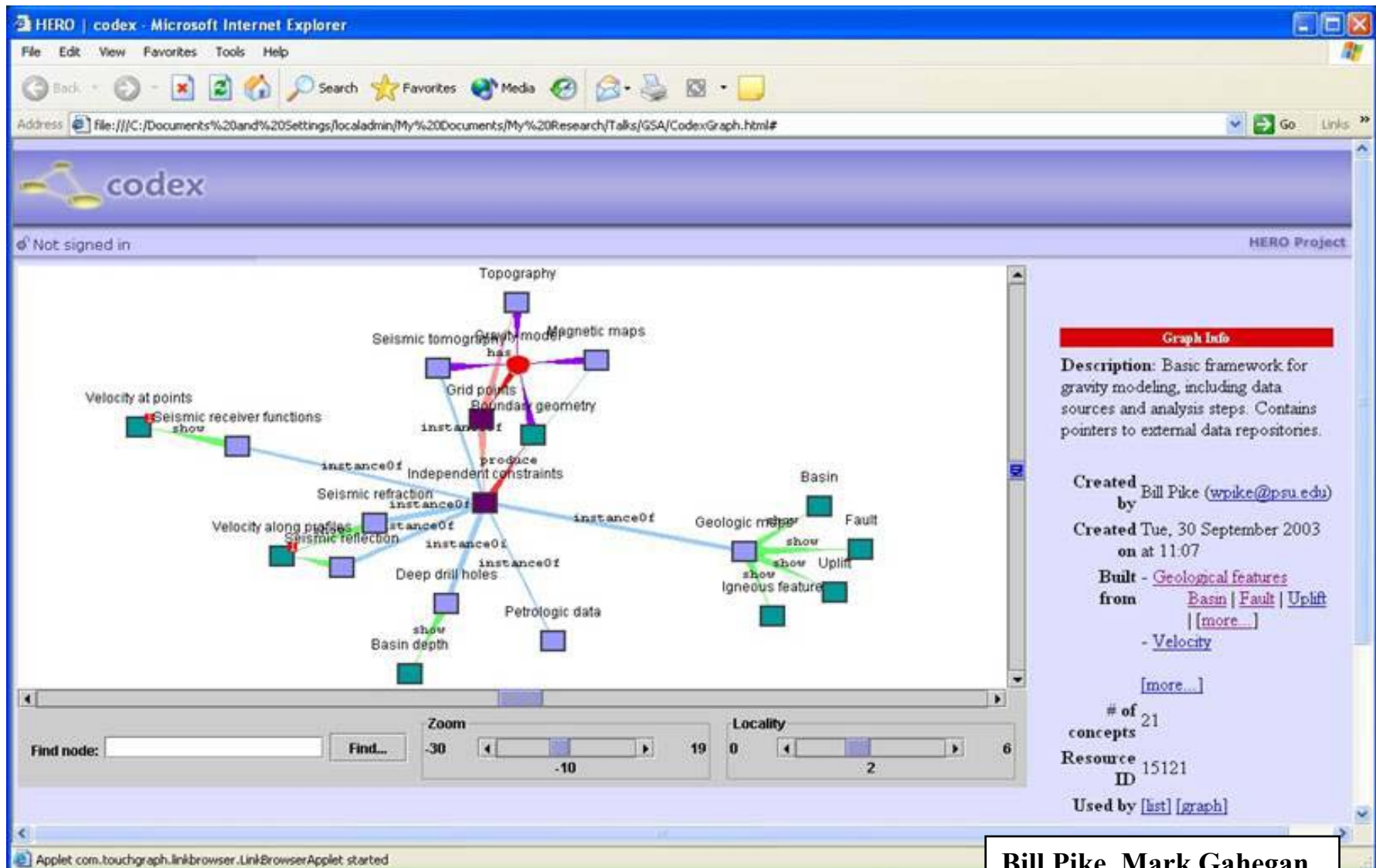
 my codex

© 2003 HERO and The Pennsylvania State University, except as noted.

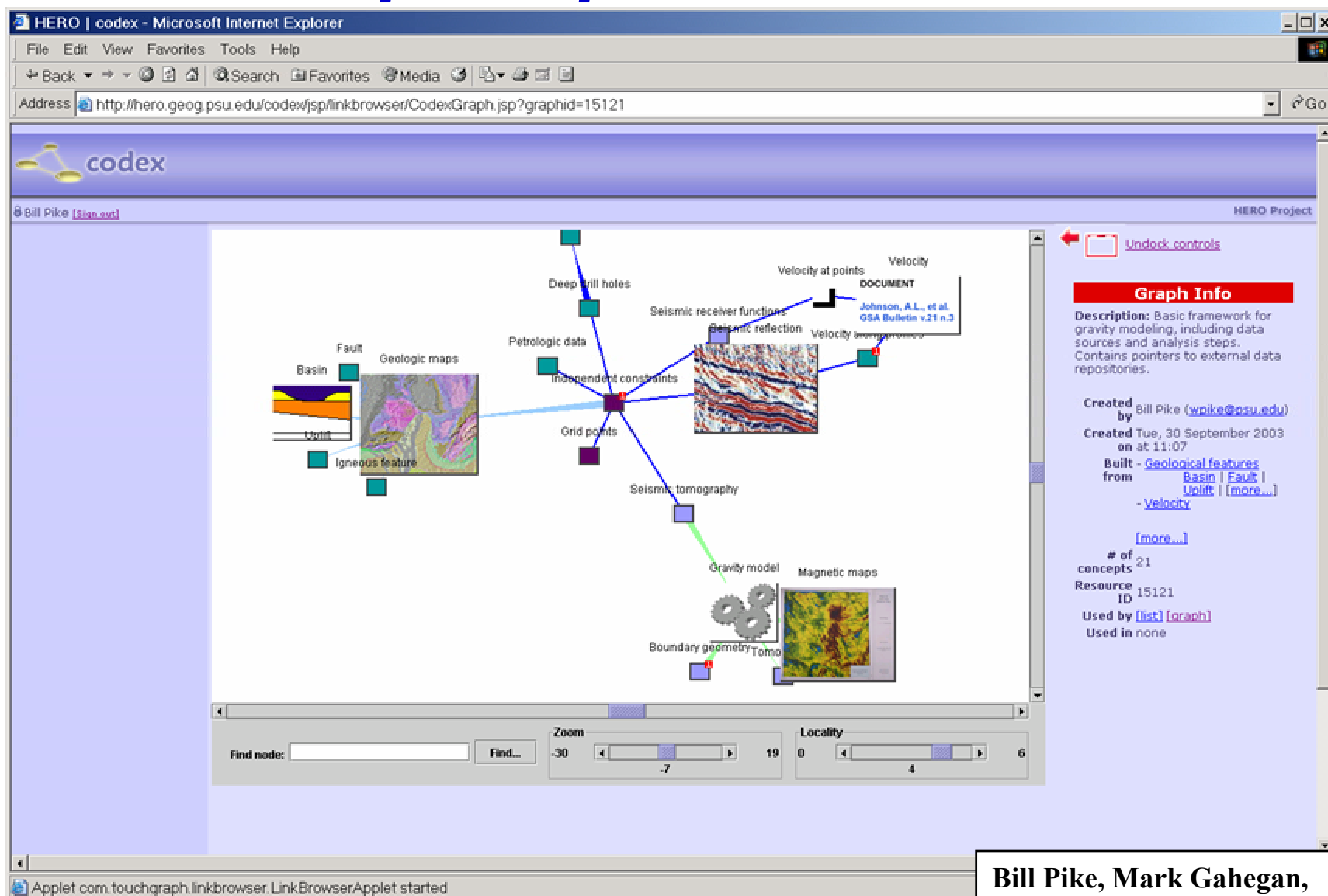
Bill Pike, Mark Gahegan,
Penn State

Concept maps: (Randy Keller's gravity map)



Bill Pike, Mark Gahegan,
Penn State

Concept maps... extend to data



Bill Pike, Mark Gahegan,
Penn State

...and to people, situations, methods

HERO | codex - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Search Favorites Media Print

Address <http://hero.geog.psu.edu/codex/jsp/linkbrowser/CodexGraph.jsp?graphid=15122> Go

codex

Bill Pike [\[info\]](#) HERO Project

Graph Info
Viewing usage info for concept **Gravity Model**

Description: Basic framework for gravity modeling, including data sources and analysis steps. Contains pointers to external data repositories.

Created by: Bill Pike (wpike@psu.edu)
Created on: Tue, 30 September 2003 at 11:07

Built from: [Geological features](#)
[Basin](#) | [Fault](#) | [Uplift](#) | [\[more...\]](#)
[- Velocity](#)

[\[more...\]](#)

of concepts: 21
Resource ID: 15121
Used by: [\[list\]](#) [\[graph\]](#)
Used in: none

James
Experimental Petrology team
Junyan
Rockies group
Gravity model
Isaac
Jamison
Alistair

Find node: Find...
Zoom: -30 19
Locality: 0 6

Applet.com.touchgraph.linkbrowser.LinkBrowserApplet started

Bill Pike, Mark Gahegan,
Penn State

Collaboration “Modes”

- **Before:**
 - Collaborate on standards, etc. that will help bring science resources online, e.g. development of schema and ontology standards
- **During:**
 - Collaborate by jointly using online resources, and “doing the science”, e.g. online analysis and mining of geologic databases, or other data sets
- **After:**
 - After doing a large computational run/experiment, or series of runs, collaborate to analyze the results, e.g. analysis of earthquake simulation runs

Challenges

- **Moving from individual PI-oriented research to collaborative research (or from individual dept./agency to inter-agency)**
 - How to deal with “re-purposing” of data and information?
- **Incentives for sharing and cooperation**
- **The “Field of Dreams” – “*If you build it, they will come*”**
 - Will you build it so that they will come, or
 - Will they come, and then you will build it
- **Also, need robust, stable, easy to use tools and environment**

SDSC/Cal-(IT)² Synthesis Center

- **Vision**

- To facilitate interactions and sharing ideas among scientists from multiple disciplines and sub-disciplines to solve *multi-disciplinary* and *multi-scale* science and engineering problems in a *collaborative* way
- To use cyberinfrastructure as a *facilitator* for the next generation of science

- **Joint activity at UC San Diego**

- Between SDSC and California Institute for Telecommunications and Information Technology (Cal-(IT)²)



Synthesis Center

- **Physical location where collaborators come together to run experiments and study experimental results using cyberinfrastructure tools**
- **Environment with ...**
 - Large-scale, wall-sized displays
 - Links to on-demand cluster computer systems
 - Access to networks of databases and digital libraries
 - State-of-the art data analysis and mining tools
- **Linked, “smart” conference rooms between SDSC and Cal-(IT)² buildings on UCSD campus**

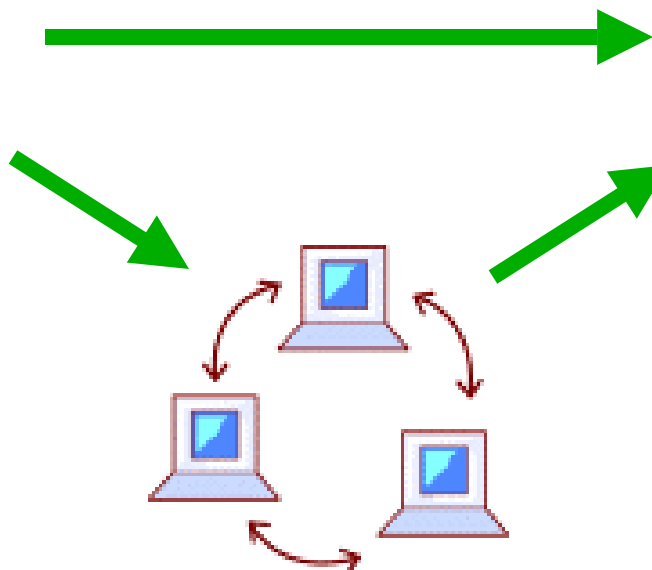


Synthesis Center

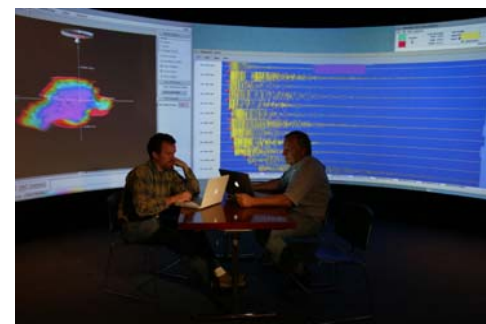
- Using the Synthesis Center



Collaboration to set up experiments



Collaboration to run experiments



Collaboration to study experimental results

The SDSC Notebook

PI: Greg Quinn, Synthesis Center, SDSC

A desktop application to better enable the scientific researcher and knowledge worker utilize network information resources and manage data

Feature List

- Leverages features of Windows and the .Net development paradigm
- Local db with search functionality
- “Knowledge” of data types
- Ability to annotate stored data
- Peer-to-peer querying of stored data and annotations
- Data export capability to popular formats
- Unattended/automatic data updates via background use of web services & HTTP
- User notification of new data
- Plug-in API for data visualization components – c/w basic data viewers for popular Bio-data types, e.g. text, protein sequences, molecules etc.
- Smart client framework for SOAP-based, data-intensive, web services
- Point-and-click interface to support new breed of Tablet PC’s and ink data types



Acknowledgements

- **Greg Quinn, PI and Team Lead**
- **Blair Jennings, Software Lead**
- **Bob Byrnes, Application Developer**
- **Mark Miller, Project Consultant**
- **Dan Fay & Microsoft Research**

<http://www.notebookproject.org>

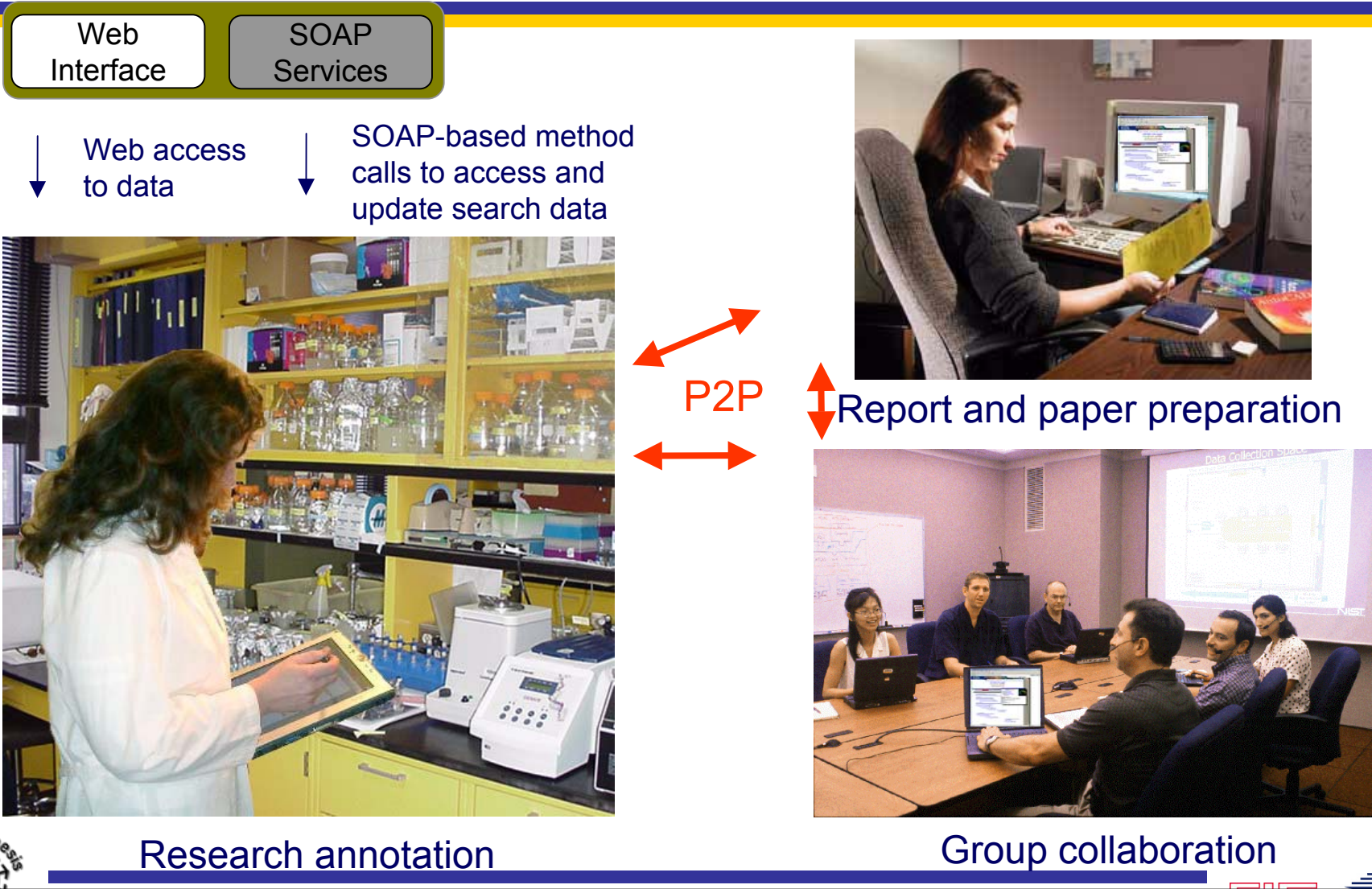


The SDSC Notebook

- **Personal data repository**
- **Smart client for web services**
- **Advanced data presentation & annotation options**
- **Collaboration environment**
- **Scheduled automated data updating**

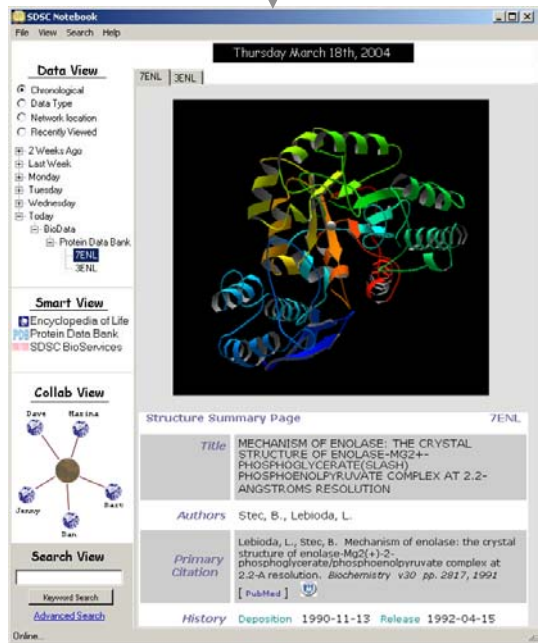


Connected research environment



Networked Data Source

Local XML data store sharable by
P2P SOAP-based communication



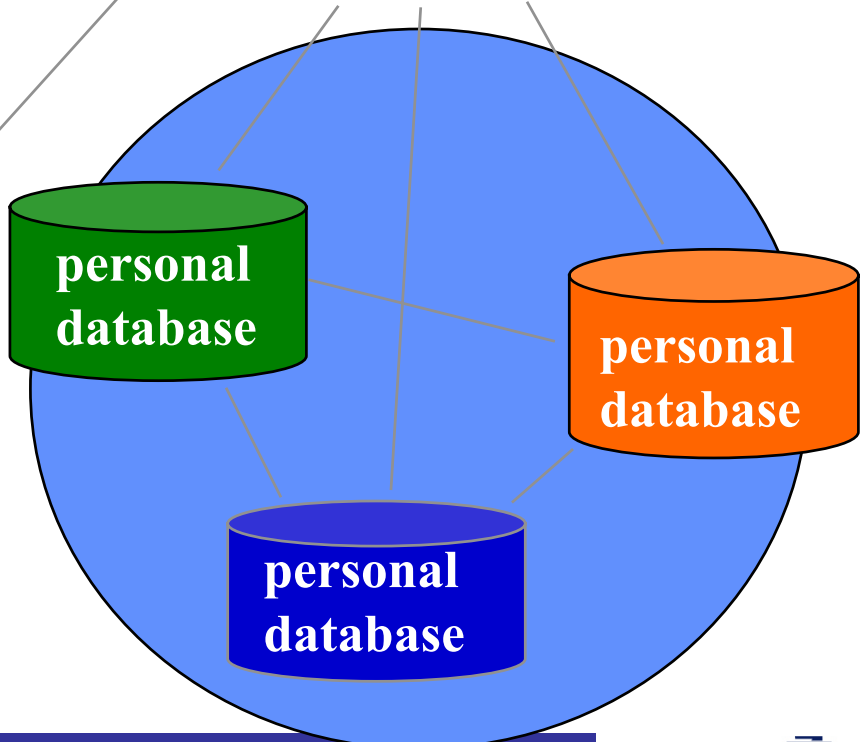
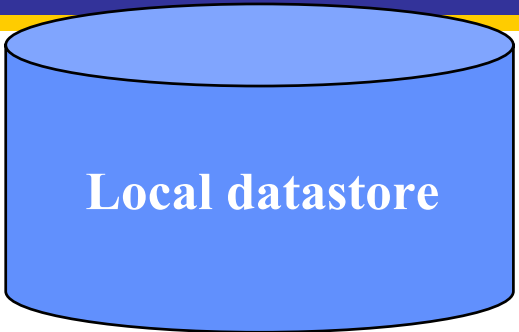
Data presentation and Smart
client for network data services



Toolbar to support web data integration

XML doc

XML doc



Prototype design of the Notebook Application

Data browser

Smart client
availability

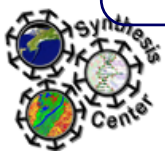
P2P
Collaboration
group

Fast search
options

The screenshot shows the SDSC Notebook application window. The title bar reads "SDSC Notebook" with menu options "File View Search Help". The date "Thursday March 18th, 2004" is displayed. The interface is divided into several sections:

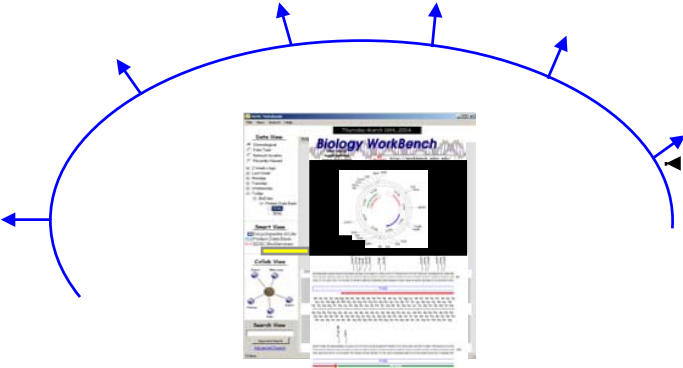
- Data View:** Includes radio buttons for "Chronological", "Data Type", "Network location", and "Recently Viewed". Below are expandable tree views for "2 Weeks Ago", "Last Week", "Monday", "Tuesday", "Wednesday", "Today", "BioData", and "Protein Data Bank". The "Protein Data Bank" section shows "7ENL" and "3ENL" selected.
- Smart View:** Lists "Encyclopedia of Life", "PDB Protein Data Bank", and "SDSC BioServices".
- Collab View:** A network diagram with a central node and five peripheral nodes labeled "Dave", "Marina", "Jenny", "Dan", and "Bart".
- Search View:** Contains a "Keyword Search" text box and a link to "Advanced Search".
- Structure Summary Page:** Displays details for entry "7ENL".
 - Title:** MECHANISM OF ENOLASE: THE CRYSTAL STRUCTURE OF ENOLASE-MG2+-PHOSPHOGLYCERATE(SLASH) PHOSPHOENOLPYRUVATE COMPLEX AT 2.2-ANGSTROMS RESOLUTION
 - Authors:** Stec, B., Lebioda, L.
 - Primary Citation:** Lebioda, L., Stec, B. Mechanism of enolase: the crystal structure of enolase-Mg2(+)-2-phosphoglycerate/phosphoenolpyruvate complex at 2.2-A resolution. *Biochemistry* v30 pp. 2817, 1991 [PubMed]
 - History:** Deposition 1990-11-13 Release 1992-04-15

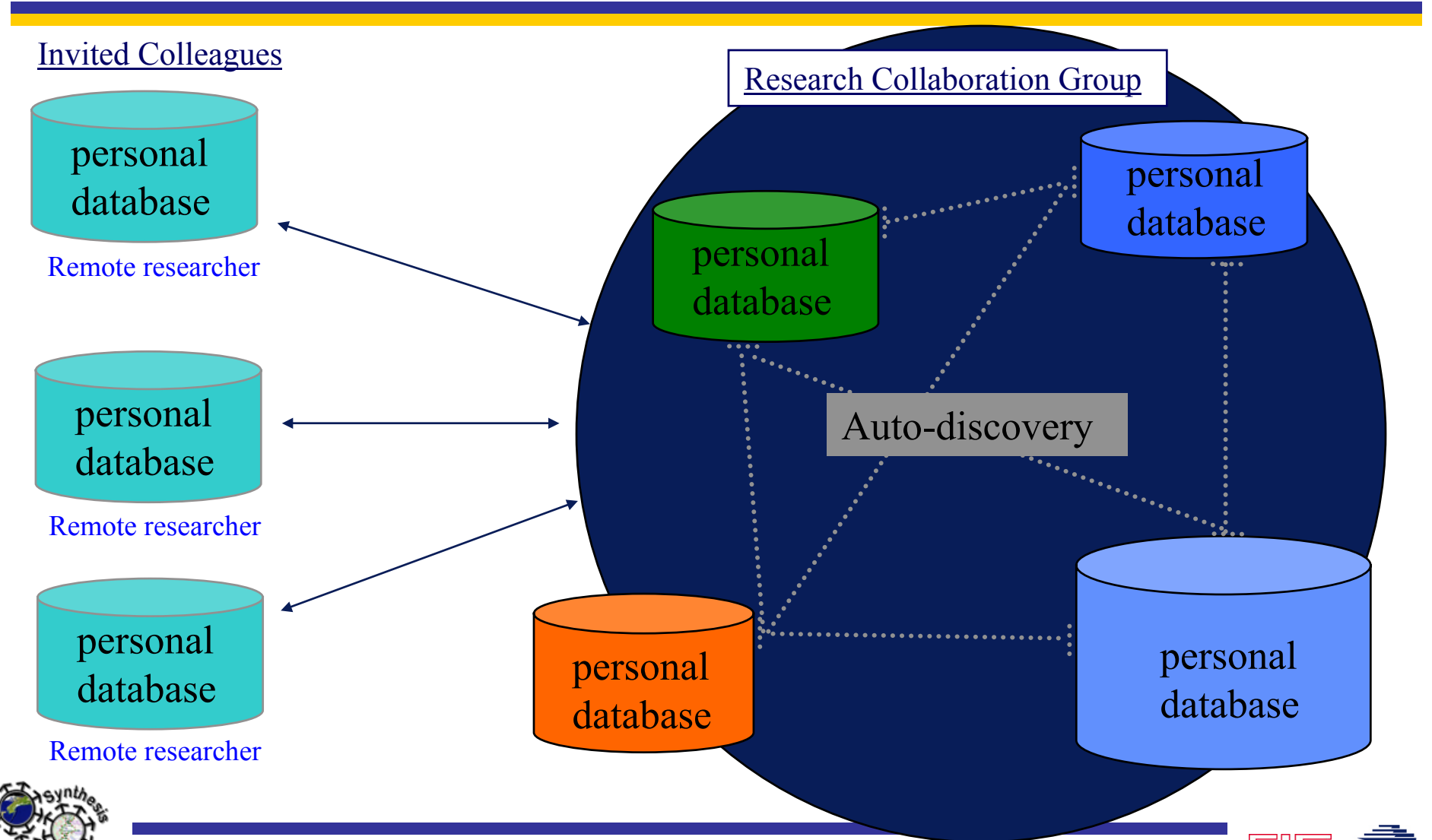
Data display
area



Data sharing

Research Collaboration Group





Data Sharing & Sociological Issues/Compliance

- Data sharing initiatives have a high priority (e.g. with NIH)
- Likely in the PI's interest that there be complete data sharing amongst her/his researchers internally and limited data sharing with external collaborators
- **But...**
- In many labs, postdocs are highly competitive and are unlikely to want to share everything
- Data needs to be tagged to indicate whether it can be shared or is invisible to others within a research collaboration group.



Alpha/Beta Testing Program

- **Identify suitable labs to partner with in software testing**
- **We will develop data viz components and advanced interfaces to data and analytical services to meet their needs**
- **We will provision new sources with SOAP-based data services where needed**
- **Garner feedback from labs, make appropriate changes to software, publish results and make software publicly available**



iGEON – International Cooperation

- **Approach: Need a geoscience and/or IT rationale for collaboration**
- **Canada**
 - Host datasets via Web Mapping Service (WMS) Server at Geological Survey of Canada, Vancouver, BC
- **China**
 - Computational Geodynamics Lab will host a GEON cluster for iGEON in China. Will work on parallelization of codes.
- **Australia**
 - Link with their AEON effort (Earth and Ocean Network)
 - Work with Dietmar Mueller to help run mantle convection codes on Linux clusters and provide as a Web service in GEON
- **Mexico**
 - CICESE (Ensenada) will host data sets on server connected via high speed network.
- **UK**
 - e-Science Center will host a GEON node at Edinburgh

For Further Information

- **Contact: Chaitan Baru, baru@sdsc.edu**